Author1, F., Author2, S., and Author3, T. (2025). Transcribing manuscripts in mensural notation: the Alamire dataset and ALOMURE pipeline, , V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

TECHNICAL SUMMARY

Transcribing manuscripts in mensural notation: the Alamire dataset and ALOMURE pipeline

Yoeri Uyttendaele[*], Bart De Moor[†]

Abstract

We present a complete Agnostic Optical Music Recognition pipeline for transcribing music in mensural notation: starting from a scanned handwritten or printed manuscript, our model produces a MusicXML or MEI file, from which the user can preserve the mensural notation or transcribe to Common Music Notation (CMN). As our second contribution, we introduce the *Alamire* dataset written in mensural notation comprising annotated folia from illuminated choir books, printed works and chansonniers. This dataset aims to be representative for polyphonic renaissance music written in monophonic notation: folia span the whole of the renaissance period, contain several hand writing styles and show various degrees of image degradation. On this dataset and the SEILS dataset, our *ALOMURE* (Alamire Optical MUsic REcognition) pipeline achieves state-of-the art music symbol recall. Our pipeline builds on an ensemble of YOLOv7 object detection models with which we mitigate class imbalance and the problem of small symbol detection. Moreover, we construct a novel staff line parsing algorithm that detects all lines on the constituting staff. Our staff line detection algorithm is capable of consistently detecting even severely degraded staff lines, and is equally applicable to the domain of music written in polyphonic notation. With our OMR pipeline, we facilitate access to music written in mensural notation for early music researchers, musicians and their audiences. With our dataset we aim to fill the gap of a versatile benchmark dataset for mensural notation. In doing so, we provide a means to make validation of OMR research on mensural notation even more robust.

Keywords: Computer vision, Optical Music Recognition, Data set

1. introduction

In past years, libraries and musicological institutions around the world have made great strides in digitizing their collections of music manuscripts (2). The field of Optical Music Recognition (OMR) aims to extract the musical information from a scan of a musical score and write this information to a machine-readable format.

In this paper we present a full pipeline OMR for music written in mensural notation: provided a scan of a printed or hand-written score, we detect the music symbols, optionally correct mistakes and write the music information to the MEI format (35) or **mens (36), the Humdrum format for mensural notation. With the aid of *Mensural-score-to-CMN-score*, a toolkit to translate mensural MEI files into common music notation, a transcription to Common Music Notation is possible (38), after which one can produce the associated audio waveform of the piece at hand. Alternatively, one can edit the output of our model online with specialized tools such as (22) or the Measuring Polyphony editor (21).

This pipeline is designed to be the applicable to the whole of western Renaissance polyphony: a large variety of handwriting styles, degraded documents, choir books, chansonniers and printed works. Moreover, we demonstrate techniques to overcome previously existent and general OMR challenges such as class imbalance and small symbol detection. Additionally, we introduce a new staff detection paradigm surmounting previous staff line retrieval issues on degraded staffs (39).

The remainder of this paper is organized as follows: Section 2 describes the state of optical music recognition as applied to early music while Section 3 outlines our dataset. In Section 4 we propose and detail our OMR pipeline. Section 5 presents the results of our proposed method on the test data. Lastly, Section 6

[*]ESAT-STADIUS KU Leuven
[†]ESAT-STADIUS KU Leuven

contains a brief conclusion of our work.

2. Related work

(37) (34) (9) (14) (15) The ultimate goal in the field of Optical Music Recognition is to create a fully automated system capable of capturing the musical semantics as written in a digital music sheet. Unfortunately, only a handful of OMR systems exist for early music manuscripts. Perhaps the most well-known is *Aruspix*(31), Aruspix produces a musical score based solely on a manuscript scan, and comes with a graphical user interface in which the user can edit the result alongside the original manuscript. Torras et al. aim to alleviate the amount of editing by using a language model (40). Pugin et al. have compared Aruspix to another system named *Gamut*, but concluded that Gamut is inferior to Aruspix, as it neither produces a musical score, nor does it come pre-trained for early music (32). Huang et al. produce a pipeline for early music notation, however they overfit their model on a single handwriting style and deploy a staff parsing algorithm that displays difficulty on degraded staffs (19). The work of Huang et al. builds upon the *classical* OMR system architecture, as specified by Rebelo et al. (33). This architecture consists of multiple subsequent steps, in different works of research these steps are modified or merged. For example, in 2018, Hajič and Pecina have shown how a convolutional neural network (CNN) can perform the *isolating primitive elements* and the *symbol classification* steps (33) of the OMR pipeline, but their method still requires a separate *binarization* step (17). Although binarization is indeed an important step in the OMR pipeline by Rebelo et al., it is far more common for CNNs to work with (the original) full-color images: using full color images as input instead of grayscale images provides the network with more information. Castellanos et al. have shown that fully convolutional neural networks can separate foreground from background in manuscript scans (12).

For early music manuscripts, separating foreground from background information is a particularly pertinent and cumbersome task: music symbols can be faded, or bleed-through from the other side of the page. Several authors have proposed methods on how to handle these issues, including Burgoyne et al. (7), who compare and combine models which take into account the image on the backside of the folium, and models that do not. Furthermore, the output of the binarization step can be used to locate the region of interest on musical imagery, even when the folium's 3D orientation in a photograph needs to be taken into account (4). In the same work, it is made clear how the location of staffs can be inferred from the binarized image by using histogram analysis. In this case, the staffs are located based on the location of *any* musical content, rather than looking for the staffs themselves. In (15) Fornes et al. introduce a ground truth database for staff lines, these lines are subject to a variety of distortions such as curvature or line interruption; in (14) the problem of finding this staff lines was addressed. Pacha and Calvo-Zaragoza attempt the entire music symbol recognition step using a single model for early music (28). The authors use a second convolutional neural network to classify the position of a detected note on the staff, in order to determine its pitch. Although Pacha et al. focus on a specific mensural writing style in their work, the underlying characteristics of the dataset are similar to those of this work: the authors attempt to detect a large number of classes (32) using the same model, while the prevalence of these classes is highly imbalanced. This work too is trained on but a single corpus. Pacha and Calvo-Zaragoza mention that for the least prevalent classes – "dots, barlines, or all types of rests" – the recall drops heavily. Huang et al. also mention issues with dot symbols and attribute this to class imbalance in the training dataset (19). In (34) Rizo et al. introduce MuRET, a closed source tool for symbol recognition, transcription and and encoding. The focus of MuRET is on two specific repertoires, monodic handwritten scores containing traditional Spanish music and white mensural notation from the 16th to 18th centuries from handwritten manuscripts. One of these reference data sets for mensural notation containing early Spanish music is the Capitan data set (9).

3. Dataset

We introduce and publish our training, validation and testing dataset, a dataset that is designed to fill the gap of representative datasets for renaissance music written in mensural notation. Our dataset contains annotations to scans of illuminated folia: from choirbooks, to printed works and chansonniers. In total we present 251 annotated folia spread across eighteen books. The images show a variety of handwriting styles at different levels of degradation and ink bleedthrough.

The dataset comprises pages from 18 manuscripts, ten of these manuscripts are illuminated choirbooks:

1. *Brussels, Koninklijke Bibliotheek van België / Bibliothèque royale de Belgique Ms. 228, 38 pages*
2. *Brussels, Koninklijke Bibliotheek van België / Bibliothèque royale de Belgique Ms. IV.922, 46 pages*
3. *Brussels, Koninklijke Bibliotheek van België / Bibliothèque royale de Belgique Ms. 11239, 18 pages*
4. *'s-Hertogenbosch, Illustre Lieve Vrouwe Broederschap Ms. 72A, 52 pages*
5. *Città del Vaticano, Biblioteca Apostolica Vaticana Ms. Chigi C VIII 234, 44 pages*
6. *Città del Vaticano, Biblioteca Apostolica Vaticana, Capp.Sist.15, 16 pages*
7. *Jena - Thüringer Universitäts- und Landesbibliothek Ms. 4, 20 pages*

8. *Munich, Bayerische Staatsbibliothek D-Mbs Mus. Ms. F, 28 pages*
9. *Vienna, Österreichische Nationalbibliothek, Musiksammlung Mus. Hs. (A-Wn) 15495, 6 pages*
10. *Città del Vaticano, Biblioteca Apostolica Vaticana S.Maria.Magg.32, 28 pages*

The codices in the above list are accessible via the IDEM database website (2). All pages are digitally photographed instead of scanned to avoid damaging the historic documents. In the above list, the first five codices originate from the workshop of a single scribe, Petrus Alamire. Although two voices may appear on the same page, the voices are annotated separately, in particular they are not aligned. This is illustrated in Figure 1.

The database also contains five chansonniers, i.e. books that are handwritten and generally contain fewer ornamentation and more degradation than choirbooks:

1. *Copenhagen, The Royal Library, MS Thott 291 8°, 20 pages* (13)
2. *Dijon, Bibliothèque Municipale, Ms. 517, 12 pages* (13)
3. *Wolfenbüttel, Herzog August Bibliothek, Codex Guelf. 287 Extravag., 6 pages* (13)
4. *Nivelle, Paris, Bibliothèque nationale, Rés. Vmc. ms. 57, 10 pages* (13)
5. *Leuven - Alamire Foundation B-AF Ms. 1, 58 pages* (2)

Finally, our database contains folia from three printed works:

1. *Ghirlanda de Madrigali a quatro vocci, 30 pages* (International Music Score Library Project)
2. *Madregali ariosi a quatro voci, 28 pages* (gar)
3. *Bologna, Compieta, Falsi Bordoni, Mottetti, Et Litanie Della Madonna A Sei Voci., Belli, 40 pages* (Giulio Belli)

The listed individual folia are available for downloaded from (41). A list of the page distribution is shown in Table 1.

The ground truth is created by an expert in Early Music Notation by manually labeling 60 pages from these manuscripts, these labels are created in bounding box format with a standard annotation program, (24). Next, we train a preliminary convolutional neural network on these labeled pages. This trained model labels another 191 pages, subsequently the expert revises each symbol and corrects the errors to obtain a dataset of 251 pages and 56962 music symbols. An annotation contains both a symbol label and a staff position label. Aforementioned data is published along with this work, with each staff annotated separately in both the CSV and **mens formats. The CSV file contains the bounding box coordinates of the music symbols as well as the symbol type and staff position.

The folia contain 57 classes that fall into eight categories, namely: *notes* (longa, breve, semibreve, minim, semiminim, fusa, semifusa), *colored notes* (longa, breve, semibreve), *ligature notes rests* (longa, breve, semibreve, minim, fusa), *clefs* (c clef, f clef, g clef), *accidentals* (flat, sharp), *mensuration signs* (Modus

book	pages	staffs	symbols
A-Wn 15495	3	22	603
NL sHerAB Ms 72A	26	236	5956
Jena Ms4	10	110	4338
Maria. Magg 32	14	116	3325
Mus. Ms. F	14	117	3195
BBr Ms 228	19	145	4579
Leuven Chansonnier	30	135	3899
BBr Ms IV 922	23	192	5250
BBr Ms 11239	9	55	1917
Capp. Sist 15	8	51	1365
Chigi Codex	22	160	5988
Dijon Chansonnier	6	12	464
Copenhagen Chansonnier	10	46	1526
Wolfenbüttel Chansonnier	3	14	3830
Gardano Primo Ariosi	14	75	2953
Anerio Ghirlanda Sacra	15	110	3531
Belli Compieta 1607	20	136	3315
Nivelle Chansonnier	5	32	928

Table 1: Overview of the pages, staffs and music symbols annotated per book in the Alamire data set.

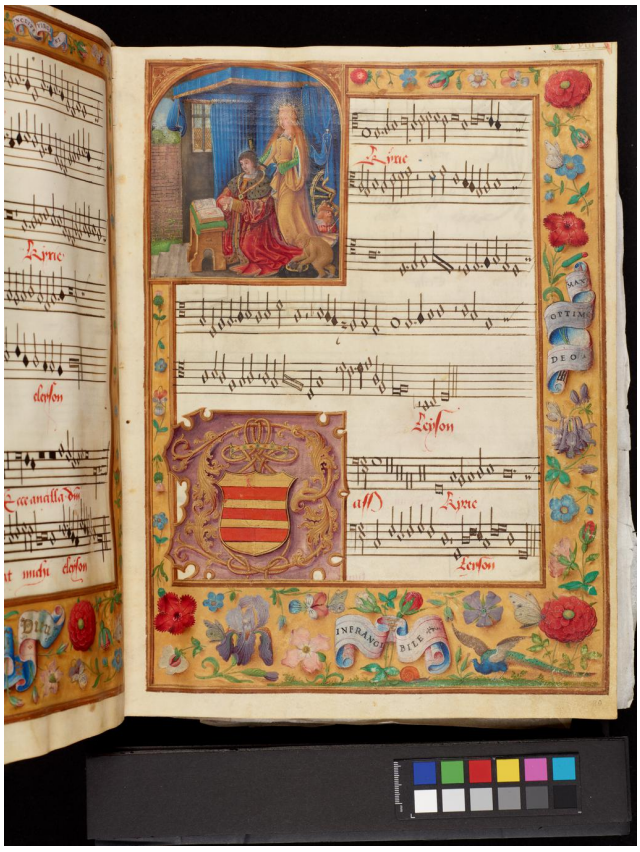


Figure 1: Folium 20r from the Chigi codex containing an excerpt from the *Missa Ecce ancilla domini* by Johannes Ockeghem. The folium has eloborate illuminations as well as heraldry (2).

(im)perfectum tempus, prolatio (im)perfectum major/minor, tactus brevis/semibrevis), *digits* (1, 2, 3),

and *others* (repeat, barline, fermata, dot, custos, congruence).

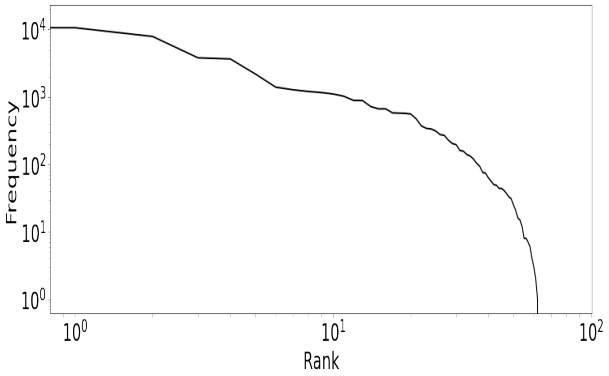
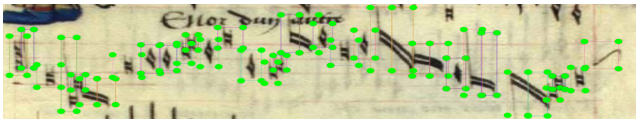
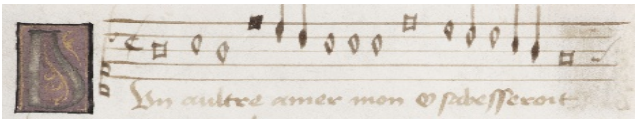


Figure 2: The rank-frequency distribution of the music symbols in our dataset. We note a pronounced class-imbalance, e.g. minims are orders of magnitude more common the the perfect mensuration sign or the *1* digit.

In the ground truth, we annotate obliques or any other type of ligature note by putting an *o* in front in the case of obliques and an *l* in the case of ligatures. Furthermore, for the sake of training and parsing, we add a 1 if the ligature note occurs as the first element in the cluster and a 2 if not. Next, we add the duratation value of note as parsed according to Apels heuristic (3). We do not interpret the written notes, e.g. we do not change duration for coloration or distinguish dots of division or augmentation. Other than ligature notes, music symbols are annotated solely based on their geometry. In other words other than the application of Apel's heuristic, we employ the agnostic notation. The dataset is made public and can be accessed via (41).

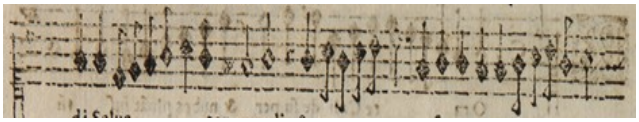


Handwriting style from the Copenhagen chansonier.



Handwriting style from the Leuven chansonier.

Handwriting style from BBR Ms. 228.



Printing style from Ghirlanda de Madrigali a 4 voici.

Figure 3:
Samples from our dataset containing chansonniers, hand written choirbooks and printed works. The works are predominantly written in white mensural notation, but feature colored symbols as well. Works vary from being in pristine condition to having missing notes and bleedthrough symbols.

4. Methodology

In this section we elaborate our OMR pipeline, in the evaluation section we further argue its components. Our pipeline differs from the classical one as proposed by Rebelo et al. (33): the border removal step is now largely superfluous given the staff detection by the affiliated specialized neural network, while the binarization step is absolved by the symbol detection neural network. Indeed, removing the staff may lead to the unwanted removal of music symbols, therefore we choose to keep the staff intact. The staff is detected by a state-of-the-art neural network, YOLOv7(42). Subsequently, we segment the found staff and apply specialized models to determine the symbol class. The segmentation allows us to overcome the small size of symbols such as *dot* or *fusa rest*. Moreover these neural nets are trained by undersampling the most frequent classes and using super classes for the rarest of symbols. Additionally, we train a network to detect bounding boxes around isolated text be it words, syllables or letters. Probabilistic filtering does away with low confident false positive staffs and false positive symbols. Detection of the lyrics allows us to filter text confounded with music symbols. A siamese network compares rare symbols with a database and untangles the *fusa-semifusa* and mensuration signs. The inter staff line spacing is estimated on the basis of the size of the music symbols and staff bounding box height. Next, a dynamic programming approach infers the staff line position after which a CovNet (23) assigns a staff position to the music symbols when applicable. Next, we parse the rest super classes and apply the ligature heuristic by Apel(3). We write away the found symbols to several formats, one is the PascalVOC XML standard upon which the output can visualized and corrective steps can be made by uploading the XML file to an annotation program such as (24). The other output formats are **mens and MEI; after conversion to these formats the user can choose to deploy the toolkit of M. Thomae to convert the mensural notation to Common Music Notation. The resulting pipeline is visualized in Figure 4.

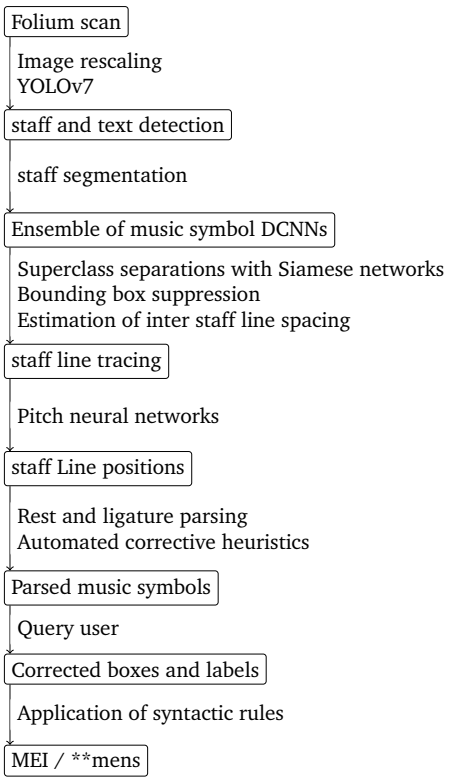


Figure 4: Starting from a folium scan, we obtain a transcribed music document. Specialized models aim to increases the duration character error rate while a staff parsing algorithms designed to handle degraded staffs aims at minimizing the positional character error rate.

4.1 staff and text detection

Scans or pictures may contain various forms of noise surrounding the folium of interest such as rulers, color checks or parts of adjacent pages. However, we choose not to implement automated border cropping techniques or staff removal techniques, as these – much like staff removal techniques – may lead to the removal of true positive symbols (8). For this same reason, in the context of music symbol detection, we abstain from binarization prior to object detection. To distinguish foreground from background, we let the CNN learn from degraded images, thereby alleviating us from the need of a binarization step (27). For the purpose of detecting the staffs, we rescale the input image to 1024×1024 pixels. After the staffs are detected, we filter false positive staffs by requiring the staffs have at least 80 % of their width within the left and right margin. These margins are themselves inferred from the staff detection output: the left margin is defined as the 20th percentile of the x_{min} coordinates of the staff box coordinates found by the neural network, analogously the right margin is the 80th percentile of the x_{max} coordinates of the staff box coordinates. If a staff has sufficient overlap with the previously defined margins but does not wholly overlap, it is nonetheless not trimmed to fit the median margins. Upon obtaining the staff boxes we estimate the median staff height by observing that, as a result of tight bounding box annotation, the predicted staff height is fairly constant at $\frac{N_{l-1}}{N_{l+1}}\Delta_l$ with N_l, the number of staff lines and Δ_l the inter staff line spacing. With the staff height, albeit crudely, estimated, we resize the image such that the inter staff line spacing is approximately 20 pixels. This number is an upper bound chosen such that the staff line inference algorithm is no longer sensitive to small perturbations relative to the inter staff line spacing, see the section on staff line tracing later on. For all object detection purposes i.e. staff, text and music symbol object detection, we choose the YOLO architecture as its Feature Pyramid Network and anchors are designed to facilitate the recognition of objects at different scales, an advantage over models with fixed anchor box sizes. Specifically, we use transfer learning on the baseline YOLOv7 model that is pre-trained on the MS COCO data set (43; 26). To to test the generalizability of our pipeline, we use k-fold cross validation, where $k = 6$ in our case, with each fold containing three unseen books from the Alamire data set. Prior to training the YOLOv7 networks, we use vertical and horizontal data augmentation whenever applicable, we disable every other default form of data-augmentation as provided by (43; 26).

Table 1 provides an overview of the unaugmented statistics of each fold. The text model is ancillary with respect to staff and symbol detection: although for later research we want to detect and parse lyrics, our current objective is to simply detect isolated words and syllables with the aim of preventing false positive lines and music symbols. For the text model, we preproccess the image by slicing it in ten equally sized crops along its height, the surface overlap between two crop is fixed at 25%. This image partitioning approach is used for training as well as inference. The text and staff models are trained using the SGD optimizer and weight decay. We use early stopping regularization. We grid search the (communal) validation set of the first five partitions to find that a confidence level of 25 % results in the least amount of false negatives. We discuss the removal of these false positives later on.

4.2 Music symbol detection

As discussed previously, music notation has inherent class imbalance, moreover, symbols such as dots or rests are small. Next to using the YOLO architecture, we mitigate these problems by performing object detection on a segmentation of each staff with an ensemble of specialized models. The segmentation is such that the staff segment width W is equal to the staff height H, the surface overlap of segments is set to 25%. In the evaluation section we compare several such ratios and we find $\frac{W}{H} = 1$ to be the optimal one. staff segmentation allows us to mitigate class imbalance: prior

Super class	Constituent symbols
rests	longa rest, breve rest, semibreve rest, minim rest, fusa rest
perfect tempus	minor prolation, major prolation
imperfect tempus	minor prolation, major prolation
flag symbols	fusa, semifusa

Table 2: The very rare symbols in this table have a geometrically similar counterpart and are distinguished by means of a Siamese network. To the object detection network these geometrically similar symbols are seen as belonging to a super classes.

model type	Constituent symbols
rare symbols	longa, fusa,l1 breve, l2 breve, l1 semibreve, longa rest, o2 semibreve, f clef , g clef, flat, barline,breve rest, semibreve rest, fermata, custos, minim rest, imincut, o1, o2, l1 l2 semibreve,o1semibreve,l2,o1semibreve,o2semibreve, sharp
ligatures symbols	l1, l2 , o1 , o2, colored l1, coloredl2 , colored o1 , colored o2
begin of staff	imaj, pmaj, imin, pmin, imincut, pmincut 3, 2, 1, c clef, f clef, g clef, flat
end of staff	custos, longa, fermata, barline

Table 3: A single staff is divided in excerpts. Specialized, fine-tuned detection models are trained to mitigate the class imbalance of the distribution in Figure 2. Each model provides its detection outputs on a given segment upon which the outputs of different models are collected.

to training we withhold excerpts featuring no rare symbols from certain categories hence attenuating the imbalance of the distribution as seen in Figure 2. Moreover, by segmenting the staffs we feed the deep convolutional neural network a dimensionally normalized and zoomed-in crop, the latter facilitating small symbol detection. Each segment is resized to 640×640 pixels, during training as well as inference.
Next to our baseline model, which is trained on all symbols in a given cross validation partition, we create specialized models with the aim of sampling the distribution in Figure 2 in a more balanced way. Our approach is two-fold: on the one hand, we introduce super classes, classes containing symbols with geometrically similar features, i.e. symbols prone to be mutually confounded by an object detection network that does aim to distinguish all possible symbols, see Table 2. On the other hand we construct a series of specialized models sampling the rank-frequency distribution in a more equilibrized manner, see Table 3.

4.2.1 Specialized deep convolutional neural networks
To construct the specialized models, we segment the staff prior to training and demand that at least one symbol in the crop belongs to the specialized training set, if not, the segment is discarded. Furthermore, if all symbols in a segment display horizontal or vertical symmetry, we augment the crop data accordingly. As is the case for the staff and text detection neural networks, we choose the YOLOv7 architecture for object detection and we randomly apply data augmentation techniques emulating degradation: image downscaling and subsequent upscaling, blurring, lowering image contrast, saturation and brightness. All models are trained and inferenced on segments of dimension 640×640. No classes were discarded in the specialized models i.e. labels outside of the classes in the specialization are still fed to the network during training. We initialize the weights of the specialized models by using the learned weights of the model trained on all symbols (the non-specialized model). The objective function is optimized with stochastic gradient descent (SGD), training is halted in accordance with early stopping regularization.

The non-specialized model and the rare model inference on every staff segment. The begin-of-staff and end-of-staff model are deployed only in the first two and two segments respectively.

4.2.2 Disentangling super classes

In order to distinguish major and minor prolation in mensuration signs or fusas from semifusas in our dataset, a fully supervised neural network would not work quite as well as a consequence of the relative sparsity of these symbols. Instead we train a Siamese neural network, a one shot learning model to distinguish similar and disparate elements. The architecture of the Siamese network (6) we constructed is shown in Table 4. The network is trained on all elements in the fold upon which the model weights are fine-tuned on the super classes. We use binary cross entropy loss and update the weights with SGD, we deploy early stopping regularization. Like all ResNet18 based networks mentioned in this text, training is done using PyTorch (30). The label with the greatest mean similarity across the elements in train and validation set is chosen as the untangled super class label.

The outputs of the ligature model are super classes as displayed in Table 3. In order to parse them according to Apels ligature heuristic, we need to detect whether: the ligature note is first in the cluster, last, or in the middle, furthermore we need to detect the color of the element, relative position (ascending/descending) and lastly, the stem orientation. The first two parameters can be derived from the output of the neural network, we discuss inference of staff line position later on. To detect the stem orientation, we detect vertical lines present in the bounding box of the ligature note by Gaussian blurring the image with a kernel size equal to the minimal stem length. The latter defined by the length of the line between the end of the stem and the height position of the symbol on the staff and is taken to be 4/3 the inter staff line spacing. Next we deploy adaptive thresholding using the inter staff line spacing as the kernel size and setting the thresholding constant $C = 6$, a value found through grid searching. Subsequently, we find the edges of the stem with canny edge detection(10) and the probabilistic Hough line transform (25), both as implemented in the

OpenCV library(5). The maximum line gap is set to 1/4 of the inter staff line spacing. If multiple lines are detected we choose the orientation of the longest line, furthermore a stem line is required to have an angle between 60 ° and 120 °.
The *rest* super class is disentangled by a ResNet neural network (18) with 18 layers in a similar vein to how positions are detected on the staff; we refer the reader to Section 4.5 for more details.

4.3 Box suppression in monophonic notation

We distinguish boxes which generally have a high degree of overlap, namely *fermata, custos, dot* from those which do not. The boxes belonging to these two categories are filtered with Non Maximum Suppression setting the allowed intersection over union to no more than 50 % and the confidence level of the retained symbols to be no lower than two standard deviations from the mean as taken over all symbols appearing in the folium. After applying the aforementioned procedure to the highly overlapping symbol category and *other* category are then fused together. Lastly, if a music symbol box has an intersection over union of more than 50 % with a text bounding box, the music symbol is discarded.

4.4 Staff line detection

Staff line detection is of primordial importance to derive note pitches. Unfortunately, staffs can be skewed and have constituent staff lines degraded to the point of invisibility, moreover staffs can overlap with lyrics. Because of the image degradation we are forced to piece together staff line evidence from across the staff, as generic CNNs are not ideal for non-local detection tasks and require ample ground truth, we choose to design a specialized algorithm instead. Another advantage of this non-end-to-end method is that we can normalize position detection by centering vertically around each symbol, see the subsequent section.
Broadly, the procedure we employ to each staff of a given folium is the following: 1. We crudely estimate the inter staff line height - but not the position- using the bounding boxes of the symbols 2. Using this estimate we partition the staff along its width into fragments with a fixed width and a height spanning the space between the upper and lower margin of the staff. We detect the staff lines in each segment of the partition. Each line in the detected histogram then gives rise to a 5-line staff proposition which is scored on the number of lines found in the histogram. We accumulate the scores of the middles of the inferred staffs as we slide across the staff width. The middle with the maximum score is the estimated middle. 3. From this found middle and the histograms of detected lines in the staff partition, we reconstruct the staff autoregressively and subject to inter staff line spacing constraints. In the following sections we detail the aforementioned procedure.

4.4.1 *Detecting staff lines in a local staff excerpt*

When detecting staff lines in a staff excerpt E of variable width W, we require the excerpt height H_s to contain N_l staff lines, where in the case of mensural notation we set $N_l = 5$. To binarize the staff excerpt, we first prepare a rectangular structuring element with dimensions $\Delta_L \times W_{se}$, with Δ_L the estimated inter staff line spacing and W_{se} the width of the structuring element. To retrieve the horizontal lines in the excerpt, we apply this structuring element as a kernel for the morphological dilation operator. Next, we use adaptive thresholding with a Gaussian kernel, i.e. the threshold value $T(x, y)$ is the cross-correlation with a Gaussian window of the $\Delta_L \times \Delta_L$ neighborhood of (x, y) minus C, where C is the adaptive thresholding constant. As degradation and bleed through may vary from excerpt to excerpt, C and W_{se} are bespoke to the given excerpt E and are determined by:

$$\underset{C\in\mathbb{N}\cap[4,14],W_{se}\in\{1,2\}}{\operatorname{argmin}} \{|L(C, W_{se}) - N_l|\}. \tag{1}$$

Where $L(C, W_{se})$ is the number of lines found through histogram analysis in the binarized excerpt, given the hyper parameters C, W_{se}. This is to say, for each staff excerpt we grid search the adaptive threshold and horizontal structure width to obtain the best approximate of five staff lines thereby minimizing noise while simultaneously mitigating image degradation. The previous heuristic is built upon the implementation of the dilation operator and adaptive thresholding as found in the OpenCV library(5).

4.4.2 *Inter staff line spacing determination*

The previous heuristic detects staff lines given a prior estimate of the inter staff line spacing. To acquire this estimate, we utilize the output of the object detection neural network: bounding boxes and their corresponding object labels from the symbol deep convolutional neural network and the box containing the whole of the staff as obtained by the staff deep convolutional neural network (DCNN). Specifically, let h be the bounding box height and f_h the expected ratio of the bounding box height with respect to the staff height then we observe that:

$$\Delta_L \approx \frac{h}{(N_l - 1) f_h}. \tag{2}$$

In the above approximation we have partitioned the the box height of the music symbols into three categories adhering to observational evidence for the parameter f_h: one category with $f_h = 1$, containing (semi)minims, (semi)fusas, rests and longas, a category with $f_h = 1/2$, with breves and semibreves and a final category of other symbols with symbols of a more variable height to which we assign $f_h = 6/4$, the estimated height of the bounding box containing the staff. The previous estimation of f_h does rely on our annotation of rests being such that they span the whole staff

and that –with few exceptions– the box drawn around the staff was about 50 percent larger than the height of the staff i.e. the group of 5 staff lines. Each of the symbol boxes with width w then induces a staff excerpt E with dimensions $H_s \times w$ where H_s spans the height between the upper staff margin and lower staff margin. By Equation 2 every bounding box has an estimate of the inter staff line spacing, $\widetilde{\Delta}_L$, this initial estimate is then used to find staff lines in the induced histogram in accordance with the heuristic described in the previous section. Let Δ_i be the difference in the height coordinates of the two subsequent lines l_i and l_{i+1} in the histogram of E, then in order to discern true spacings, we subject Δ_i to:

$$\left| \frac{\widetilde{\Delta}_L - \Delta_i}{\widetilde{\Delta}_L} \right| \leq \epsilon, \tag{3}$$

where $\epsilon := 1/3$ is the allowed variability as a fraction of the initially estimated inter staff line spacing. Finally, the inter staff line spacing Δ_L for a given staff is then given by the median of the sampled valid spacings of all bounding boxes.

4.4.3 Estimating the middle of the staff

Staffs can be skewed and/or degraded, the skewness of the staff lines prevents us from using a global estimate of the staff middle: the larger the staff gradient the smaller the window is where a constant middle estimate is applicable. On the other hand, degradation compels us to have an adequately large sample of the staff in order to find enough staff line evidence.

First, we partition the staff in segments along the staff width. The sampling frequency f is set to $f = \frac{1}{3}\Delta_L$, a parameter found through grid searching allowing us to not introduce too much noise in the subsequent binarization process while still being capable of handling staff skewness. The dimensions of a staff segment are given by $\Delta_L \times H_s$. For every staff excerpt we detect the lines in the histogram, the idea is then, to given this data, identify which lines belong to the staff. If text appears in the segment we truncate the height of the box to prevent false positive staff lines in the form of text. A staff line necessarily appears as a part of a group of 5 somewhat regularly spaced staff lines, this observation allows us to eliminate lines which are impossibly far removed from what are assumed to be true staff lines, or conversely, we can fill-in missing data so long as the result is a group of five regularly spaced lines. Thus for each segment E we want to obtain $\max_{\text{5-lines}} \Pr(\text{5-lines}|\text{data}) = \max_{\text{middle}} \Pr(\text{middle}|\text{data})$, where the equality trivially follows from the observation that the middle of the staff dictates the position of the other staff lines given a regular spacing. We now define the set of possible middles associated with the lines observed in E. Let $\{l_i\}$ be the set of lines found in the histogram through our binarization procedure, then naively, line l_i could be identified with each of the possible five lines s_j of the true staff, from bottom to top staff line. Each one such identification induces a staff estimate and therefore a middle estimate m_{ij}. Every m_{ij} is attributed a score, reflecting the evidence for the 5 staff lines k_{ij} around it:

$$\begin{cases} \sigma(m_{ij}) & = \sum_{k_{ij}}^{N_l} \mathbb{1}_{evidence}(k_{ij}), \\ \sigma(m_{ij}) & = 0 \quad \text{if } k_{ij} \text{ not within margins.} \end{cases} \tag{4}$$

In Equation 4 a line is considered in the evidence if it appears in histogram of the segment E. From this definition it follows that the maximum middle score in a noiseless segment with all middles found would be $\sigma = 25$, the lines adjacent would be attributed a score $\sigma = 16$, with the outer lines having a score of $\sigma = 9$. As a corollary from the previous discussion, we limit the maximum scores to $\sigma = 25$. Furthermore two middles m_1 and m_2 are identified provided they have the same score and $|m_1 - m_2| < \epsilon$. With this we posit that $\Pr(\text{m}|\text{data}) \propto \sigma(m)$.

If a staff has no noise and no staff lines are faded, then the middle line of the staff could be extracted from but this one segment E. Unfortunately, the previous conditions aren't always realized and we therefore use the global staff information. As stated, each segment in our partition gives rise to middle estimates along with their respective scores. The middle estimate for the staff would then be the middle with the highest total score. The height position of the middle of the staff may vary along its width, therefore, we need a way to identify and update the middle as we slide across the staff width – in our case from left to right. Let m_i be a middle estimate of staff excerpt E_i and let $\{m_j\}$ be the identifiable middles -within a distance smaller than ϵ- to m_i from the adjacent excerpt E_{i+1}. Denote their respective scores by $\{s_j\}$, then we trace the continuation m_{i+1} of the line containing m_i as

$$m_{i+1} = \frac{\sigma_i m_i + \sum_j \sigma_j m_j}{\sigma_i \sigma_i + \sum_j \sigma_j}. \tag{5}$$

$$\sigma_{i+1} = \frac{\sigma_i \sigma_i + \sum_j \sigma_j \sigma_j}{\sigma_i \sigma_i + \sum_j s_j \sigma_j}. \tag{6}$$

with σ_{i+1} the score of m_{i+1}. Finally, the middle line M is then given by $M = \max_m \sigma(m)$.

4.4.4 Full staff reconstruction

The staff is reconstructed by propagating the middle along the staff width and updating the positions of the constituent lines in accordance with the staff line evidence in each sample:

$$l_{m+1} = \frac{1}{a} \sum_{n=1}^{a} l_{m+n}, \tag{7}$$

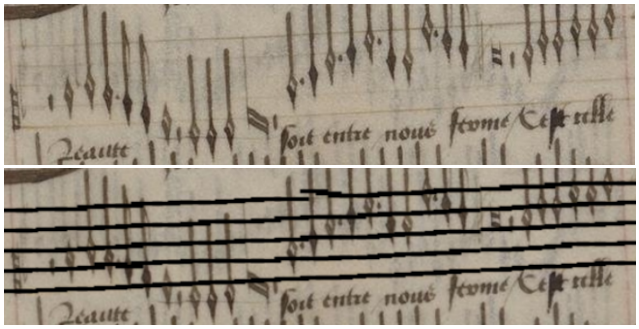


Figure 5: A degraded and skewed staff from Folium 16v of the Dijon Chansonnier. The staff is degraded, skewed and has text directly underneath it, all these factors hinder staff line detection and tracing.

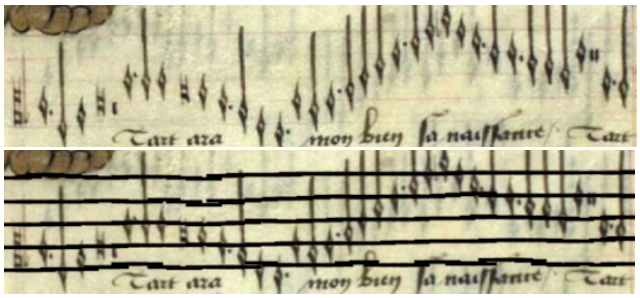


Figure 6: A degraded staff from page 15 from the Copenhagen chansonnier. The staff is severely degraded, particularly in its first half thus determining the lines of the staff requires global staff information. Moreover, we require the filtering of text such that the latter isn't confounded with the lower staff line.

with $|l_m - l_{m+n}| < \epsilon$ and hyper parameter $a = 5$, corresponding to the amount of samples we look ahead in determining l_{m+1}. This parameter is found through grid searching. In case of severe degradation where no lines are found the line position is propagated to the next sample. The full staff is defined to contain all the N_l drawn lines and in addition two ledger lines above and 2 below the group, regardless of whether they are drawn and extrapolated on the basis of the inter staff line spacing.

4.5 staff line position detection

Knowing the coordinates of all the staff lines around a given symbol, we are capable of detecting the position of that music symbol on the staff. To this end, we create a staff excerpt with a width equal to the width of the bounding box, starting vertically with the upper most ledger line and ending vertically with the lower most ledger ledger line. Thus the excerpt's vertical coordinates overlap with those of the full staff line box as described in the previous section. To prevent erroneous labels as a consequence of emblems or text, we set every pixel above and below the bounding box of symbol to white, see the rightmost image in Figure 7. We rescale the image to 224×224 pixels, set it to grayscale and feed it to a ResNet18 network. Because of the static rescaling dimensions and fixed amount of staff lines in image, the positions of the staff lines become quasi normalized. Save for the case of considerable inter staff line spacing variation, it is no longer necessary to have the staff lines drawn, we can afford for them be faded and even erased in the denoising procedure.

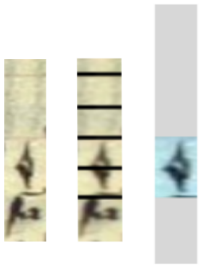


Figure 7: Attributing a position of a note on the staff, excerpt from the Copenhagen chansonnier. In the left image we note severe staff line degradation and the presence of text, factors hindering staff position detection. In the middle figure we see the model having detected all staff lines. In the right picture we have extrapolated all the lines from the middle to contain ledger lines as well, moreover we set all pixels outside of the bounding box to white.

The training data is generated by inferring the staff lines on the ground truth data, extracting the image described above and using the labeled staff position on the staff. To train the staff line detection and rest disentangling, we again use a cross entropy loss and update the weights with SGD, regularizing them with early stopping regularization. We note that in some chansonniers there's extensive variation of the inter staff line spacing of the ledger lines in comparison to non ledger lines. Therefore, starting from the the non-leger line model, we fine-tune a model specific for ledger lines. In addition, clefs positions coincide strictly with line positions, i.e. there's no line clef with a position in-between two staff lines. This allows us to narrow down the outcome space of the clef positions; we fine-tune a model specifically for clefs.

As staff lines may be degraded to the point of invisibility it is not always obvious for a neural network to distinguish between semibreve rests and minim rests. Moreover, given that rests are small and relatively featureless, their separation is more susceptible to noise. In other words, we treat the separation of the rest classes in the same way as we detect staff line positions: we first detect the staff line positions, find the associated box and denoise the picture. Here too we deploy a Resnet18 network with hyper parameters equal

to the ones used in staff position detection. Upon inference we perform horizontal and vertical data augmentation.

4.6 Post detection heuristics

Inevitably the pipeline will display errors, however, some of these errors are against (musico)logical rules. We implement a series of heuristics to remedy these logical inconsistencies.

1. Correcting shared symbols between staffs. A note can only belong to one staff, yet in the case of densely written staffs, notes can be doubly detected. In such case, the occurrence with the lowest confidence is removed from the folium detections.
2. No dots at staff line level. Predominantly in the case of printed works, staff lines can have a granular appearance such that the network confounds these staff line segments with dots. Thus, we filter dots coinciding with the staff line.
3. Fixing isolated ligatures. Sometimes the network detects isolated ligature notes, assuming there are no adjacent missed ligature note, the current ligature note must be misclassified. In this reasoning it is assumed that the oblique symbol consists of two notes. Therefore, we train a ResNet18 network trained on all symbols in the dataset save for ligature notes to attribute the true label to the symbol. The network uses weights pre-trained on ImageNet as presented in Pytorch. The weights are optimized with SGD.
4.Inserting missing clefs. In mensural notation, a clef is often present as the first symbol on the left off the staff, yet degradation they may be confounded for another symbol. We use the classification model described in the previous paragraph to go over the first three symbols on the staff, if a clef label is predicted by the classification network of the previous paragraph, we swap the label predicted by the YOLOv7 network with the predicted clef symbol.
5. perfect prolation correction. Semibreves are sometimes written as near circles therefore resembling perfect minor prolation. Therefore, the perfect minor prolation is swapped for the semibreve label if not occuring near a clef.
6. Correcting the position of flats. Flats adjacent to a clef always belong to a *si/b*. Therefore, the position of the flat on the staff–given the nature of the clef and its position– is constrained. We correct the predicted positions of the flats to align with the constrained positions.
7. Correcting the position of the custos. The custos symbol indicates the position of the the next symbol on the staff given the clef of the next staff. Given that there's considerable variation in the way a custos is written, the detection of the pitch is somewhat less reliable than that of the average note. Therefore, we use the predicted position of the next note to label the custos position.

	Component type
Backbones	2 × Resnet18
Metric	L_1 Functional Layer
Forward layers	1D Batch Normalization
	Fully Connected Layer
Activation	Sigmoid

Table 4: The architecture of the Siamese neural network used to separate geometrically similar figures. We use the L_1 distance on the 16-dimensional output feature layers of the ResNet18 networks. Training is performed with Binary Cross Entropy loss.

4.7 Musical interpretation

In mensural notation, the note duration is determined by the duration of adjacent notes, coloration, and dot of augmentation and division. We append the series of heuristics implemented by Thomae in order to convert the set of detected musical objects in mensural notation to Common Music Notation (38).

5. Evaluation and Experiments

In this section we evaluate the OMR pipeline and perform ablation studies to establish the utility of its individual components. We review the effects of crop ratios, super classes, specialized DCNNs, siamsese networks and heuristics. We evaluate our work on the SEILS dataset and by $k = 6$-fold cross validation on the Alamire dataset. We find our model compares favorably with other works. Given the need to correct the output of an OMR system for false positives and missing true positives to produce a faithful transcription, we use the normalized editing distance, i.e. the character error rate (CER) as the evaluation metric. The metric is applied to a folium as a whole and is defined in Equation 8:

$$CER = \frac{I + D + S}{N}. \tag{8}$$

Where I is the number of insertions, D the number of deletions and S the number of substitutions made in producing a correct staff with N symbols.
First, we determine the optimal aspect ratio (width-height ratio) of the window sliding over the staff. Too large a window and class imbalance will be more prominent, too small window sizes and the training of neural networks will be less stable due to remnants of other symbols on the edges and a lacking wider context for gauging ink bleed trough. In table 5 we observe the optimal durational character error rate (DCER) is obtained when the width of the staff crop is equal to its height. The test set data is that of partition 1, i.e. containing folia from the choirbooks Br Ms 11239, Maria. Magg 32 and the Dijon Chansonnier. The evaluation is that of a single model fine-tuned on rare classes.

aspect ratio	DCER
0.50	8.99%
0.75	7.74%
1.00	**6.48%**
1.25	7.03%
1.50	7.71%

Table 5: The duration character error rate for several aspect ratios as evaluated on Partition 1. An height-to-width ratio of 1 strikes the optimal balance between overfitting on the majority classes and introducing remnants of over other symbols when slicing the staff into crops.

Figure 8: An excerpt from the Leuven Chansonnier, folium 21r. The folium has a low foreground-background contrast such that quite a few symbols are unrecognized. Stems of minims are faded thereby increasing the likelihood of the model confounding minims with semibreves.

Again using partition 1 as our test data set, we establish the positive effect of using specialized models for discerning symbols of the super class categories: rests, mensuration signs, (semi)fusas and ligature symbols. In Table 6 we observe the most pronounced decrease in character error rate in the ligature category, indeed, the local, convolution based neural network is unlikely to learn to parse the constituent ligature notes according to Apel's heuristic, particularly in longer sequences.

DCER	rests	rare symbols	ligatures
with specialized model	13.49%	21.52%	10.12%
without specialized model	14.29%	25.32%	16.36%

Table 6: The duration character error rate for several specialized detection models. Ligature parsing in particular benefits from an ad hoc model.

Specialized detection model	-Δ DCER
staff beginning	2.34%
staff ending	4.48%
ligature	5.19%
rare	0.7%

Table 7: The duration character error rate for several specialized detection models. The difference in durational character error rate. The difference is taken between the full model, where every super class is present, and the full model minus the specialized DCNN in consideration. The OMR pipeline benefits from every specialized DCNN.

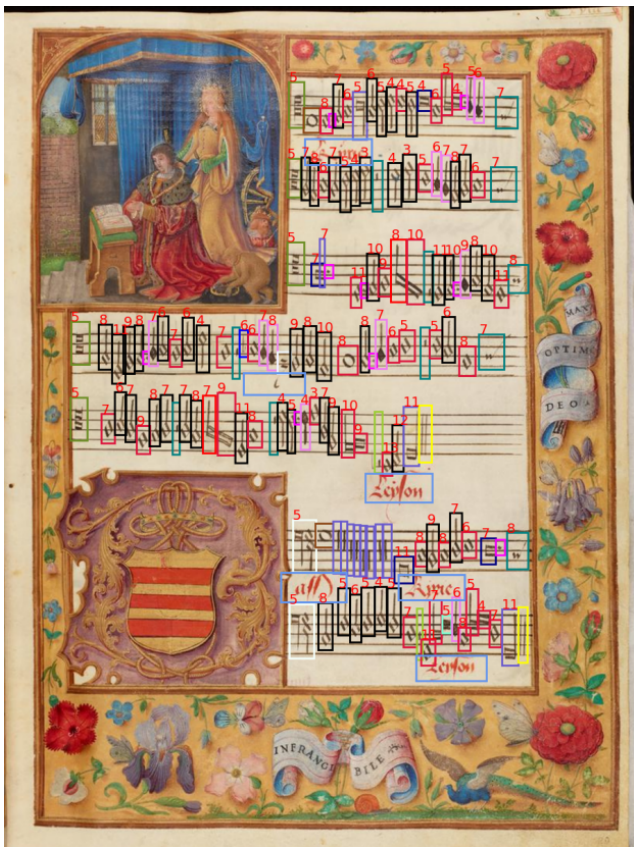


Figure 9: Object detection results on Folium 20r from the Chigi codex. The output of our model is color-coded such that the same duration value has the same color in ligatures as well. The model also detects lyrics even when written in clusters of single letters.

Heuristic number	−Δ CER
1	0.70%
2	−0.02%
3	0.05%
4	0.06%
5	0.04%
6	0.03%
7	0.04%
total	0.84%

Table 8: The impact of the heuristics on the character error rate. The impact of each the heuristics is small but cumulatively non-negligible.

We now turn to evaluating the effects of the additional specialized deep convolutional neural networks, see Table 7. The evaluation is on the first partition and each model is evaluated only on the symbols it is fine-tuned on, see Table 3. The difference in durational character error rate is taken between the *full* model, where every super class is present, and the full model minus the specialized DCNN in consideration. We note that the OMR pipeline benefits from every specialized

position detection	recall
with augmentation	96.4%
without augmentation	95.4%

Table 9: The effect of augmenting images i.e. horizontal and vertical flipping upon inference of staff positions.

DCNN.

The effect of heuristics is listed in Table 8. We note each heuristic lowers the character in an albeit marginal way save for the method to delete common symbols between staffs. We also note that the heuristic to remove dots in the vicinity of lines increases the character error rate on the given test data set. However, on printed works with granular staff lines the aforementioned heuristic does serve to lower the character error rate substantially: on the SEILS dataset the character error rate drops by 3.1 % by removing dots on on staff lines.

The effect of data augmentation i.e. the ensemble of inference on the regular image and its horizontal and vertical mirroring is shown in Table 9. We note a small but beneficial effect in symbol recall.

In Table 10 we observe the results of the 6-fold evaluation on the Alamire dataset, we refer the reader to (41) for a visualization of the results on a randomly chosen page from each book. Of the 1572 staffs in the corpus, 3 staffs were not recognized by the staff DCNN. All of these three staffs belong to the Leuven Chansonnier and are uncommonly short, containing no more than 6 symbols. Of the recognized staffs all lines were traced correctly, meaning all lines were traced correctly with deviations scarce and contained within the staff line identification bound. From Table 10 and Table 11 we note that music symbols on printed works have the lowest character error rate, 3.18%, owing to their relative uniformity, moreover the printed sources in our corpus display comparatively little ink bleed through and degradation. The Choir books display a greater variety of image quality, from the most degraded source the *Capp. Sist 15* featuring symbols illegible even to the human reader, to the more pristine *s'Hertogenbosch codex*. On the set of communal illuminated choir books, *Mus. Ms. F* and *sHerAB Ms 72A*, Van Gool et al. report an average recall of 85.6% and therefore a character error rate of at least 14.4%. On these communal codices our model achieves a character error rate of 4.01%. An example of inference on a choir book, *Chigi codex*, can be seen in Figure 9.

Lastly, the chansonniers have an average total character error rate of 7.32 %. With a 11.78% CER the Leuven chansonnier is the most difficult to parse of all works considered. Figures 11 and 8 display detection on chansonniers. Figure 11 shows how ligatures are well detected even when densely written. In Figure 8 we see how a low contrast page from the Leuven Chansonnier with faded symbols causes the model to not recognize some music symbols, in addition, stems of minims are faded such that the model confounds minims for semibreves. In general ink bleed through and fading causes confusion between white notes and their colored counter parts and symbols with stems and their non-stem analogue. Other recurring false positives are encountered with ornate letters being confused for clefs, see e.g. Figure 8 or the custos being confused with dots.

Lastly, we evaluate our work on the Symbolically Encoded Il Laurro Secc dataset (SEILS) containing 150 pages and 33 categories (29). Of the latter dataset not a single page was seen during train or validation. We perform object detection on SEILS with the model trained on the first partition, we note that the training set of this fold contains all 3 printed works (49 pages) of the Alamire dataset. Our model achieves state-of-the art results with an average CER of 3.26% on the SEILS dataset, outperforming the mean CER of 4.5% by Castellanos et al(11). The main source of errors of our model is fusa rest and semifusa recall.

fold number	Source	Position CER	Duration CER	Total CER
1	BBr Ms 11239,	3.20 ± 1.80%	3.89 ± 2.56%	3.57 ± 2.14%
	Dijon Chansonnier	6.27 ± 4.68%	7.73 ± 2.34%	7.07 ± 3.22%
	Maria. Magg 32	6.55 ± 4.25%	5.94 ± 3.32%	6.21 ± 3.63%
2	Gardano Primo Ariosi	1.48 ± 0.95%	1.99 ± 0.98%	1.76 ± 0.71%
	Anerio Ghirlanda Sacra	1.99 ± 1.16%	5.00 ± 1.38%	3.67 ± 1.04%
	BBr Ms IV 922	2.82 ± 1.93%	2.88 ± 2.97%	2.86 ± 2.38%
3	Mus. Ms. F	6.52 ± 5.14%	5.06 ± 3.80%	5.72 ± 4.34%
	Belli Compieta 1607	2.81 ± 3.18%	5.20 ± 3.16%	4.12 ± 2.90%
	BBr Ms 228	2.81 ± 1.56%	3.02 ± 1.58%	2.93 ± 1.32%
4	Capp. Sist 15	6.60 ± 6.70%	5.96 ± 6.12%	6.26 ± 6.24%
	Copenhagen Chansonnier	7.85 ± 3.81%	5.15 ± 2.31%	6.42 ± 2.78%
	NL sHerAB Ms 72A	2.43 ± 2.65%	2.20 ± 1.68%	2.30 ± 2.01%
5	Jena Ms4	3.89 ± 2.49%	3.73 ± 1.65%	3.81 ± 1.81%
	Leuven Chansonnier,	10.36 ± 7.06%	13.01 ± 7.25%	11.78 ± 6.77%
	Wolfenbüttel Chansonnier	4.15 ± 1.83%	4.27 ± 0.11%	4.21 ± 0.88%
6	Nivelle Chansonnier	9.61 ± 1.32%	9.80 ± 3.28%	9.71 ± 2.27%
	Chigi Codex	4.21 ± 2.33%	5.90 ± 3.44%	5.14 ± 2.61%
	A-Wn	2.43 ± 2.65%	2.20 ± 1.68%	2.30 ± 2.01%

Table 10: An overview of the different character error rates for the pipeline for all the 6 folds in the cross validation evaluation.

	Position CER	Duration CER	total CER
Chansonniers	7.15 ± 3.74%	7.54 ± 4.28%	7.32 ± 3.81%
Printed	2.09 ± 2.02%	4.06 ± 1.48%	3.18 ± 1.55%
Choir books	4.22 ± 3.15%	4.08 ± 2.27%	4.21 ± %
SEILS	3.18 ± 2.71%	3.32 ± 1.63%	3.27 ± 1.32%

Table 11: Summary of the recall for the ensemble model with super classes and sub-sampling. Both pitch and duration recall are high. Unsurprisingly, best results on the printed works as they show a uniform handwriting style and generally display little signs of image degradation.

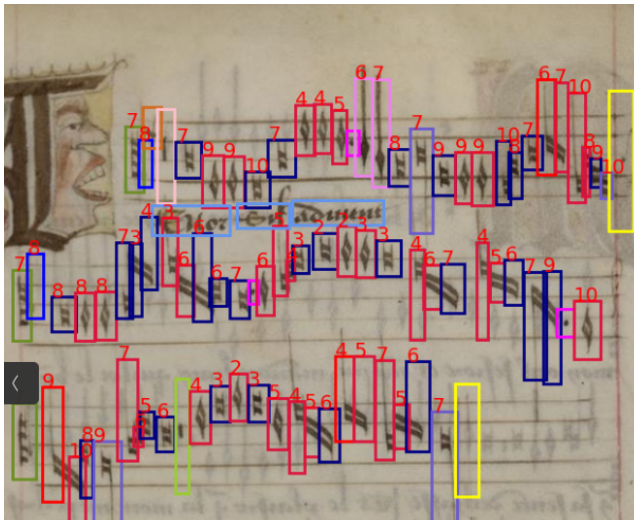


Figure 10: A crop of folium 10bis, from the Dijon Chansonnier (13). We can see that ligature notes are densely written but well recognized. The custos' are in the margin and faded and therefore missed. Similarly, a faded flat in the bottom staff goes undetected.

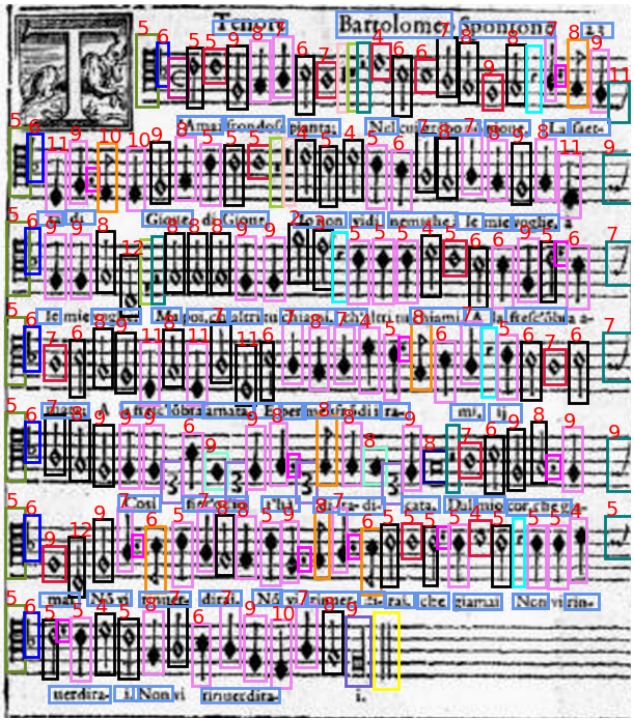


Figure 11: A page from the SEILS corpus. All music symbols are properly detected. The text model also performs well on this corpus.

6. Conclusion

In this work we demonstrated that inherent challenges of Optical Music Recognition on early documents such as image degradation, class imbalance, and small scale symbols can be successfully dealt with. Faded and missing symbols may still pose a problem. Deviating from the classical OMR pipeline by Rebelo et al, we used bounding box information to detect staff margins and uses specialized models for symbol detection. Additionally, an algorithm was devised to deal with possibly degraded and skewed staffs. Moreover, for the benefit of future OMR and musicological research, we make our annotated dataset publicly available. Lastly, the methods we proposed can be adapted to detect plainchant or modern notation. Our ALOMURE model achieves state-of-the-art symbol recall on a variety of source types and handwriting styles and by incorporating the possibility for users to rectify mistakes and automatically transcribe works to Common Music Notation, we offer researchers in polyphonic music a way to process our collective knowledge of an important part of Western cultural heritage.

In future work, we will expand our corpus to not only include text bounding boxes but also the textual content, as well as expanding and providing semantic representations of our dataset. Furthermore we will investigate digital restoration as a pre-processing step to obtain lower character error rates. Moreover, we will further investigate how musicological constraints from counterpoint can aid us in building more accurate models, in particular upon building a more easy to maintain end-to-end (vision) transformer based model. In its current form our pipeline is available as a downloadable repository for offline usage. In the future however, we hope to integrate it as web application for the IDEM website.

References

[gar] Madrigali ariosi a quattro voci composti da diversi eccellentissimi autori. Libro secondo.

[2] Alamire Foundation and Research Group Musicology KU Leuven (2021). Polyphony – IDEM Integrated Database for Early music.

[3] Apel, W. (1961). *The notation of polyphonic music, 900-1600*. Number 38. Medieval Academy of Amer.

[4] Barbancho, I., Segura, C., Tardón, L. J., and Barbancho, A. M. (2010). Automatic selection of the region of interest in ancient scores. In *Melecon 2010-2010 15th IEEE Mediterranean Electrotechnical Conference*, pages 326–331. IEEE.

[5] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

[6] Bromley, J., Bentz, J., Bottou, L., Guyon, I., Lecun, Y., Moore, C., Sackinger, E., and Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25.

[7] Burgoyne, J. A., Devaney, J., Pugin, L., and Fujinaga, I. (2008). Enhanced bleedthrough correction for early music documents with recto-verso registration. In Bello, J. P., Chew, E., and Turnbull, D., editors, *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, pages 407–412.

[8] Calvo-Zaragoza, J., Barbancho, I., Tardón, L. J., and Barbancho, A. M. (2015). Avoiding staff removal stage in optical music recognition: application to scores written in white mensural notation. *Pattern Analysis and Applications*, 18(4):933–943.

[9] Calvo-Zaragoza, J., Toselli, A. H., and Vidal, E. (2017). Handwritten music recognition for mensural notation: Formulation, data and baseline results. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1081–1086.

[10] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

[11] Castellanos, F. J., Calvo-Zaragoza, J., and Inesta, J. M. (2020). A neural approach for full-page optical music recognition of mensural documents. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, pages 23–27.

[12] Castellanos, F. J., Calvo-Zaragoza, J., Vigliensoni, G., and Fujinaga, I. (2018). Document analysis of music score images with selectional auto-encoders. In Gómez, E., Hu, X., Humphrey, E., and Benetos, E., editors, *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pages 256–263.

[13] Christoffersen, P. W. (2021). The Copenhagen Chansonnier and the 'Loire Valley' chansonniers An open access project.

[14] Fornés, A., Dutta, A., Gordo, A., and Lladós, J. (2011). The 2012 music scores competitions: Staff removal and writer identification. In Kwon, Y. and Ogier, J., editors, *Graphics Recognition. New Trends and Challenges - 9th International Workshop, GREC 2011, Seoul, Korea, September 15-16, 2011, Revised Selected Papers*, volume 7423 of *Lecture Notes in Computer Science*, pages 173–186. Springer.

[15] Fornés, A., Dutta, A., Gordo, A., and Lladós, J. (2011). Cvc-muscima: A ground truth of handwritten music score images for writer identification and staff removal. *International Journal on Document Analysis and Recognition - IJDAR*, 15:1–9.

[Giulio Belli] Giulio Belli. Compieta, Mottetti, Letanie della Madonna a otto voci. .

[17] Hajič jr., J., Dorfer, M., Widmer, G., and Pecina, P. (2018). Towards full-pipeline handwritten OMR with musical symbol detection by u-nets. In Gómez, E., Hu, X., Humphrey, E., and Benetos, E., editors, *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pages 225–232.

[18] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[19] Huang, Y., Chen, X., Beck, S., Burn, D., and Van Gool, L. (2015). Automatic handwritten mensural notation interpreter: From manuscript to MIDI performance. In Müller, M. and Wiering, F., editors, *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, pages 79–85.

[International Music Score Library Project] International Music Score Library Project. Ghirlanda de Madrigali a quatro vocci.

[21] K. Desmond, M. T. e. a. (2020). Next steps for measuring polyphony: A prototype editor for encoding mensural music.

[22] Laurent Pugin, R. Z. and Roland, P. (2014). Verovio - a library for engraving mei music notation into svg. *International Society for Music Information Retrieval*.

[23] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44.

[24] Lin, T. (2015). Labelimg.

[25] Matas, J., Galambos, C., and Kittler, J. (2000). Robust detection of lines using the progressive probabilistic hough transform. *Computer Vision and Image Understanding*, 78:119–137.

[26] MetaAI (2022). sota real-time object detection on coco.

[27] Nazaré, T. S., da Costa, G. B. P., Contato, W. A., and Ponti, M. (2017). Deep convolutional neural networks and noisy images. In Mendoza, M. and Velastin, S. A., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 22nd Iberoamerican Congress, CIARP 2017, Valparaíso, Chile, November 7-10, 2017, Proceedings*, volume 10657 of *Lecture Notes in Computer Science*, pages 416–424. Springer.

[28] Pacha, A., Calvo-Zaragoza, J., and Hajič jr., J. (2019). Learning notation graph construction for full-pipeline optical music recognition. In Flexer, A., Peeters, G., Urbano, J., and Volk, A., editors, *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 75–82.

[29] Parada-Cabaleiro, E., Batliner, A., Baird, A., and Schuller, B. W. (2017). The SEILS dataset: Symbolically encoded scores in modern-early notation for computational musicology. In Cunningham, S. J., Duan, Z., Hu, X., and Turnbull, D., editors, *Proceedings of the 18th International Society for Music In-*

formation Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017, pages 575–581.

[30] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[31] Pugin, L. (2009). Editing renaissance music: The aruspix project. In *Digitale Edition zwischen Experiment und Standardisierung*, volume 31 of *Beihefte zu editio*, pages 147–156. Walter de Gruyter, Berlin, New York.

[32] Pugin, L., Hockman, J., Burgoyne, J. A., and Fujinaga, I. (2008). Gamera versus aruspix: Two optical music recognition approaches. In Bello, J. P., Chew, E., and Turnbull, D., editors, *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, pages 419–424.

[33] Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marçal, A. R. S., Guedes, C., and Cardoso, J. S. (2012). Optical music recognition: state-of-the-art and open issues. *Int. J. Multim. Inf. Retr.*, 1(3):173–190.

[34] Rizo, D., Calvo-Zaragoza, J., and Iñesta, J. M. (2018). Muret: a music recognition, encoding, and transcription tool. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, DLfM '18, page 52–56, New York, NY, USA. Association for Computing Machinery.

[35] Roland, P. (2002). The music encoding initiative (mei).

[36] Roland, P., Hankinson, A., and Pugin, L. (2014). Early music and the music encoding initiative. *Early Music*, 42(4):605–611.

[37] Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., and Widmer, G. (2013). *Roadmap for Music Information Research*. Creative Commons BY-NC-ND 3.0 license.

[38] Thomae, M. E. (2020). Mensural-score-to-CMN-score.

[39] Timofte, R. and Gool, L. V. (2012). Automatic stave discovery for musical facsimiles. In Lee, K. M., Matsushita, Y., Rehg, J. M., and Hu, Z., editors, *Computer Vision - ACCV 2012, 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV*, volume 7727 of *Lecture Notes in Computer Science*, pages 510–523. Springer.

[40] Torras, P., Baró, A., Kang, L., and Fornés, A. (2021). On the integration of language models into sequence to sequence architectures for handwritten music recognition. In Lee, J. H., Lerch, A., Duan, Z., Nam, J., Rao, P., van Kranenburg, P., and Srinivasamurthy, A., editors, *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, pages 690–696.

[41] Uyttendaele, Y. (2022). Omr code.

[42] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.

[43] Wang, C.-Y., Yeh, I.-H., and Liao, H.-Y. M. (2021). You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*.